

## Chapter 17

# Evaluating Multi-Agent Systems

The transition from single-agent to multi-agent systems, and specifically Large Language Model (LLM)-based systems, requires a paradigm change in the evaluation paradigm. In contrast to single-agent evaluation, in which the immediate concern is performance on a particular task, evaluation of LLM-based multi-agent systems must be understood in terms of inter-agent dynamics as a whole, such as collaborative planning and communication effectiveness. Both task-oriented reasoning and holistic capability evaluation are addressed in this chapter, reflecting the nuance of such evaluations. In greater detail, there are two main areas that we examine for evaluation. First, there is task-solving Multi-Agent Systems (MAS), where we examine benchmarks assessing and enhancing LLM reasoning for coding, knowledge, and mathematical problem-solving tasks. These tests also accentuate the utility of distributed problem solving, achieved through organized workflows, specialisation among agents, iterative improvement, and calls for additional tools. Enhanced reasoning, primarily because of agent-agent decision-making cooperation and multi-round communications, is shown for MAS compared with agent-based individual ones. Following that, there is a general evaluation of MAS abilities, extending beyond one-task-oriented achievement, to agent interactions at a highly advanced level. It involves a move away from one-dimensional measurements into multi-dimensional frameworks for documenting achievements at collaborations, reasoning abilities, system efficiency, and flexibility. We categorize such measurements into collaboration-oriented and competition-oriented measurements and have identified efficiency, decision-making quality, quality of collaboration, and flexibility as primary measure domains. These measurements capture various aspects of agent behavior, including communication effectiveness, resource distribution, and response to dynamic situations.

### 17.1 Benchmarks for Specific Reasoning Tasks

In multi-agent system solving for tasks, much focus has been on leveraging multi-agent coordination for enhancing the reasoning capacity of LLMs. It is most evident in coding, knowledge, and mathematical reasoning benchmarks, where one is interested in examining and building on performance with distributed solving. These benchmarks most typically examine if agents' capability for producing correct code, reasoning on complex knowledge domains, and solving difficult mathematical problems withstanding, with measures such as *pass@k* [1076] or proof ratios for success being prevalent. Much improvement has been exhibited by MAS through structured workflow, domain-specific agent roles, and iterative improvement on state-of-the-art performance. On the contrary, for model and simulation MAS, the case is one with a comparative lack of standardized benchmarks. Rather, research is primarily experimental setups that simulate a variety of social phenomena, with calls from the community for further formalized evaluation frameworks. These multiple benchmark areas are described below, examining the tasks, measures for evaluation, and the core mechanisms through which MAS result in better performance.

**Code Reasoning Benchmark** Measuring the capability of LLMs for code synthesis requires bespoke benchmark suites with a focus on functional correctness. Code synthesis, as compared to natural language synthesis, allows for direct verification through running. Several benchmark suites have been built for this purpose, typically consisting of a collection of programming problems, each described with a natural language problem description and a collection of test cases for automatically ascertaining the synthesized code's correctness. HumanEval [1077], APPS [1078], and MBPP [939] are some popular ones. These benchmark suites predominantly utilize the *pass@k* metric, which computes the percentage at which at least one among the top-*k* generated solutions passes all test cases for a number of problems. The problems covered through these benchmark suites range across a variety of difficulties and programming

abstractions, requiring not only for LLMs and Agents but also for syntactically correct and logically sound code that satisfies the provided test cases. Recent work has explored leveraging Multi-Agent Systems (MAS) for enhancing LLM capability on code reasoning. For instance, MetaGPT [626] is a meta-programming system which embeds human-like Standard Operating Procedures (SOPs) into multi-agent cooperation based on LLM. With multi-agent role assignment with varying domains and adopting assembly line mode, MetaGPT effectively breaks down difficult operations into sub-operations and achieves state-of-the-art performance on HumanEval and MBPP benchmarks. SWE-agent [628] presents a novel Agent-Computer Interface (ACI) which largely enhances a repository-creating, repository-editing, and navigation capability for an agent. The system demonstrates that a well-structured interface tailored for LMs can largely enhance software engineering capability, with state-of-the-art on SWE-bench and HumanEval. AgentCoder [994] is another multi-agent coding system with focus on effective testing and auto-optimization. It is a three-agent system with a programmer, a test designer, and a test executor. The test designer supplies accurate and diverse test cases, and the test executor provides feedback to the programmer for optimization. Such collaborative workflow enhances coding efficiency and outperforms one-agent models and other multi-agent approaches on HumanEval and MBPP datasets. These MAS approaches all point out multi-agent cooperation, organized workflow, and tailored interface as effective solution strategies for enhancing the capability of LLM on code reasoning. DEVAI [781] proposes a set of novel AI development automation benchmarks, which utilize a judge-agent mechanism for judging automatically intermediate development process.

**Knowledge Reasoning Benchmark** To facilitate AI agents effectively acting in and understanding the world, robust knowledge reasoning abilities are essential. Benchmarks for this class assess an agent’s ability to utilize factual knowledge and logical reasoning when answering challenging queries. Commonsense reasoning is tested with benchmarks such as CSQA [1079] and StrategyQA [1080], and scientific knowledge understanding is tested with ScienceQA [1081]. The core challenge for agents is performing multi-step, chain-of-thought reasoning, stepwise logically progressing from input query to output answer. These tests concentrate on assessing how well a specific AI agent can apply a specific body of knowledge, one at a time, and reason out a problem. Recent research has experimented with the use of LLMs on MAS for improving knowledge reasoning task performance, and they have achieved state-of-the-art accuracy. For example, MASTER [1009], a novel multi-agent system, employs a novel recruitment process for agents and communication protocol using the Monte Carlo Tree Search (MCTS) algorithm, and achieves 76% accuracy on HotpotQA [940]. Reflexion [48], a universal framework for bringing reasoning and acting together with language models, improves baseline by 20% on HotpotQA. These strategies demonstrate the potential of multi-agent coordination for knowledge reasoning tasks. Besides, leveraging external tools, e.g., search engines, is also needed for improving knowledge reasoning capacity. Agents may apply these tools for retrieving the latest information and also for fact checking, thus improving the accuracy and dependability of responses. Such integration is particularly helpful on applications such as TriviaQA [1082], for which real-time information access is essential.

**Mathematical Reasoning Benchmark** Math reasoning is a critical skill for AI agents which requires cooperative utilisation of mathematical knowledge, logical deduction, and computational power. Benchmarking tasks for this capability tend to fall into two categories: math problem-solving and computer-aided theorem proving (ATP). Datasets such as SVAMP [942], GSM8K [1083], and MATH [941] challenge agents to solve word problems, asking for exact number answers or formulas. ATP is a harder test, with stricter compliance with formal proof schemata. Tests on datasets like PISA [1084] and miniF2F [1076], which are graded on proof completion, test whether an agent can produce well-formed mathematical proofs. Multi-agent systems (MAS) have been put forward as a potential solution for handling mathematical reasoning problem complexity. Methods such as MACM [1010] include a multi-agent system consisting of Thinker, Judge, and Executor agents tailored for a complex problem, dividing it into smaller sub-problems for computation. The Thinker agent generates new ideas, Judge decides if they are accurate, and Executor conducts necessary computation involving tools such as calculators. Such a modular structure supports iterative refinement and elimination of errors, enhancing problem-solving accuracy. Furthermore, methods such as multi-agent debate [985] include several instances of a language model debating and refocusing iteratively for collective solution improvement, enhancing reasoning as well as factuality accuracy. Such MAS-based systems have achieved notable improvement on benchmarks such as MATH and GSM8K, establishing distributed solving capacity for mathematical problems. Aside from this, reinforcement learning from human feedback (RLHF) and preference learning strategies have been attempted for further enhancing mathematical problem-solving capacity of LLMs. For instance, a multi-turn online iterative direct preference learning framework [1085] has been put forward for training various language models with enriched sets of prompts over GSM8K and MATH datasets. Such a technique includes feedback from interpreters for codes and optimizes preferences at a level of trajectories, with notable improvement in output.

**Societal Simulation Benchmark** Social simulation benchmarks are essential for evaluating multi-agent system performance and realism for simulating human behavior and social interactions based on LLMs. Standardized sets and test cases for evaluating the agents’ ability for interacting, communicating, and evolving within a simulated society are

Table 17.1: MAS Benchmarks: A Systematic Classification of Multi-Agent System Evaluation Frameworks Categorized by Task-Oriented Performance and System-Level Capabilities. This comprehensive collection encompasses both specialized task-solving benchmarks and holistic capability assessments, reflecting the dual nature of MAS evaluation in collaborative problem-solving and inter-agent dynamics.

Category	Focus	Benchmarks	Examples	Representative Metrics
Task-solving	Code Reasoning	APPS [1078], HumanEval [1077], MBPP [939], CodeContest [1087], MTPB [1088], DS-1000 [1089], ODEX [1090], Raconteur [1091]	MetaGPT [626], SWE-agent [628], AgentCoder [994]	Pass@k, Resolved(%)
	Knowledge Reasoning	ARC [1092], HotpotQA [940], CSQA [1079], StrategyQA [1080], BoolQ [1093], OpenBookQA [1094], WinoGrande [1095], HellaSwag [1096], SIQA [1097], PIQA [1098], proScript [1099], ScienceQA [1081], ProOntoQA [1100]	Reflexion [48], MASTER [1009]	Accuracy
	Mathematical Reasoning	MATH [941], GSM8K [1083], SVAMP [942], MultiArith [943], ASDiv [1101], MathQA [1102], AQUA-RAT [1103], MAWPS [1104], DROP [1105], NaturalProofs [1106], PISA [1084], miniF2F [1076], ProofNet [1107]	MACM [1010], Debate [985]	Accuracy, Pass@k
Collaboration	Communication-based Cooperation	InformativeBench [1108], Collab-Overcooked [944], COMMA [1109], LLM-Coordination [926]	iAgents [1108], Two-Player [1110], EAAC [1111]	Task Completion Rate Communication Efficiency
	Planning and Coordination	PARTNR [946], VillagerBench [925], BABYAGI-ARENA [1112], Multiagent Bench [948]	AAS [1113], ResearchTown [1114], GPTSwarm [651]	Planning Success Rate Coordination Efficiency
	Process-oriented	Auto-Arena [947]	Idea [1115]	Process Completion Rate Step Efficiency
Competition	Adversarial Scenarios	BattleAgentBench [920], MAgIC [955], LLMArena [1116], PokerBench [1117], Multiagent Bench [948]	Dilemma [1118], Pok��LLMon [1119]	Win Rate Elo Rating
	Social Deduction	AvalonBench [972], Human Simulacra [1120], Diplomacy [934]	MA-KTO [1121], HLR [1122]	Win Rate Accuracy of Deductions
	Game-Theoretic	Guandan [1123], AgentVerse [1124], ICP [1125]	WarAgent [1126]	Score Win Rate

provided through the benchmarks. An example of one such widely used benchmark is SOTOPIA [1086], employed for evaluating social intelligence in natural language agent-based social intelligence. It is employed for evaluating agents’ ability for conversing, understanding social cues, and building relationships with each other within a virtual society. Another benchmark involves simulating propagation Gender Discrimination and Nuclear Energy [255] topics on social networks. It is employed to evaluate agents’ capabilities in modeling opinion dynamics, information dissemination, and social influence within large-scale social networks. Multiagent Bench [948] further provides two simulation domains—werewolf and bargaining—to assess competitive interactions among diverse agent groups with conflicting goals.

Evaluating capabilities in LLM-based MAS requires specialized approaches that effectively measure the rich interactions between agents. As this field evolves, evaluation methodologies have transitioned from single-dimension metrics to multi-faceted evaluation frameworks that capture the complex skillset required for effective multi-agent interaction. This evolution reflects a growing understanding that agent performance must be assessed across multiple dimensions including collaboration success, reasoning capabilities, and system efficiency.

In recent research, the MAS evaluation can be mainly categorized along three primary dimensions: collaboration-focused benchmarks, competition-focused benchmarks, and adaptive and resilience benchmarks. Within each category, we identify specific metric domains that capture different aspects of agent performance. Current evaluation approaches typically measure efficiency metrics (e.g., task completion rates, resource utilization, time efficiency), decision quality metrics (e.g., action accuracy, strategic soundness, reasoning depth), collaboration quality metrics (e.g., communication effectiveness, coordination efficiency, workload distribution), and adaptability metrics (e.g., response to disruptions, self-correction), which provide a foundation for evaluating multi-agent systems.

**Collaboration-focused Benchmarks.** Collaboration-focused benchmarks have evolved significantly, shifting from basic single-dimensional metrics toward comprehensive frameworks that evaluate complex agent-to-agent communication

and coordination. Initial benchmarks, such as InformativeBench [1108], primarily addressed agent collaboration under conditions of information asymmetry, employing metrics like Precision and IoU to measure decision accuracy in information dissemination tasks. Subsequently, the scope of evaluation expanded, exemplified by Collab-Overcooked [944], which introduced nuanced process-oriented metrics such as Trajectory Efficiency Score (TES) and Incremental Trajectory Efficiency Score (ITES). These metrics assess detailed aspects of coordination, revealing significant shortcomings in agents' proactive planning and adaptive capabilities despite their strong task comprehension.

Further expanding the evaluation scope, COMMA [1109] and LLM-Coordination [926] emphasized communication effectiveness and strategic synchronization, employing diverse environments and extensive metrics including Success Rate, Average Mistakes, and Environment Comprehension Accuracy. These benchmarks collectively illustrate an emerging trend toward capturing deeper aspects of collaborative behaviors and strategic consistency.

Other benchmarks, such as PARTNR [946], VillagerBench [925], and BabyAGI [1112], further addressed gaps in existing evaluations by focusing explicitly on reasoning, planning, and task decomposition. These benchmarks highlighted the need for comprehensive assessment of agents' ability to engage in complex, socially embedded tasks, considering metrics like Percent Completion, Balanced Agent Utilization, and agent contribution rates. AgentBench [706], VisualAgentBench [928], and Auto-Arena [947] further standardized multi-agent evaluations, automating assessment across various domains and demonstrating substantial performance disparities between closed-source and open-source LLMs. These observations underscored critical challenges in developing universally effective collaboration frameworks.

In summary, collaboration-focused benchmarks collectively reflect an ongoing shift toward comprehensive, nuanced evaluations that encompass communication efficiency, adaptive strategy, and fine-grained agent coordination, addressing earlier limitations focused solely on outcome-based performance.

**Competition-focused Benchmarks.** Competition-focused benchmarks evaluate agents' strategic capabilities and adversarial interactions, highlighting specific deficiencies in Theory of Mind and opponent modeling. Early benchmarks such as BattleAgentBench [920] and MAgIC [955] initiated the focus on mixed cooperative-competitive environments, uncovering critical weaknesses in high-order strategic reasoning among LLM agents. These benchmarks employed comprehensive competitive metrics such as Forward Distance, Judgment Accuracy, and Rationality scores, identifying that while advanced LLMs performed adequately in simpler scenarios, significant limitations persisted under complex adversarial conditions.

Building upon these insights, subsequent benchmarks like Human Simulacra [1120], LLMArena [1116], and PokerBench [1117] further refined competitive evaluation by incorporating human-like reasoning metrics and more robust strategic measures (e.g., Response Similarity Score, Elo Scores, and Action Accuracy). These evaluations consistently demonstrated shortcomings in opponent prediction, risk assessment, and adaptive strategic planning, despite high task comprehension.

Social deduction and deception-based benchmarks, notably AvalonBench [972] and Diplomacy [934], further revealed fundamental gaps in agents' abilities to interpret hidden information and manage complex social dynamics. Metrics like Assassination Accuracy, Deduction Accuracy, and Win Rates emphasized that even sophisticated LLMs fail to replicate human-level reasoning in adversarial negotiation and hidden-information games.

Additional game-theoretic evaluations, including Guandan [1123], AgentVerse [1124], MultiAgentBench [948], and ICP [1125], introduced scenarios requiring strategic cooperation under incomplete information. These benchmarks reinforced previous findings on the necessity of enhanced Theory of Mind and predictive modeling capabilities. Multi-AgentBench [948] also introduces the KPI and coordination score to evaluate the competition of agents. Collectively, competition-focused benchmarks highlight persistent strategic and reasoning limitations among LLM-based agents, underscoring the ongoing need to address critical gaps in adversarial modeling and strategic planning despite advancements in general reasoning and task execution capabilities.

**Adaptive and Resilience Benchmarks** adaptive and resilient multi-agent system benchmarks tackle two interconnected capabilities together: adaptability—the ability of the agents to act dynamically in altering, unexpected environmental conditions by modifying their behavior and strategy. Resilience, or the ability of the system to endure, alleviate, and rapidly recover from disruptions, faults, or hostile intervention. In adaptability, as mentioned in AdaSociety [1127], the dynamic interplay between social relationships and physical environments demands that agents engage in continuous learning, and strike a balance between environment discovery and social network construction. Despite significant advancements in current multi-agent decision-making frameworks, these environments fall short in introducing new challenges in various physical contexts and changing social interdependencies. Therefore, AdaSociety introduces an environment in which physical states, tasks, and social relationships among agents continuously evolve, thereby capturing the adaptability of agents as they respond to expanding task complexity and shifting resource constraints.

Moreover, current benchmarks may oversimplify the challenges of real-world automation with limited disruption modeling and simplified dependencies of process [945], resulting in insufficient evaluation of planning capabilities and adaptability. Thus, REALM-Bench [945], on the other hand, defines adaptation through real-world-inspired planning problems, which emphasizes metrics such as real-time re-planning efficiency, coordination scalability under increasing complexity, and the stability of performance outcomes despite dynamic interdependencies or disruptive events. Conversely, resilience benchmarks [1128] systematically introduce faults or errors into individual agents to assess overall system robustness.

## 17.2 Challenge and Future Work

While various MAS evaluation benchmarks have been developed in recent years, challenges and limitations continue to exist with regard to the standardization of evaluation across different MAS tasks and scenarios, and the ability to evaluate scalability and diversity in MASs. Future research must address these challenges, in order to develop the comprehensive field of MAS evaluation.

Below are some challenges and future directions in LLM Multi-agent evaluation:

1. Multi-Agent System has demonstrated superior performance in solving complex tasks, when compared with single agent frameworks. But compared with single agent system, MAS also requires more computations and brings additional costs. Therefore, there has a urgent challenge that we need to handle: when we need to invoke MAS framework? For many simple user instructions, we may only require LLM or single agent system to accomplish. And only complex user instructions could require MAS frameworks. Hence, in the future, how to design the task router mechanism to detect which scenario require MAS or not is fundamental but also a important issue.
2. Multi-agent system is a high-level framework, built upon multiple AI agent based on the foundation models. Therefore, just like back propagation, the optimization of MAS framework will also affect each part (i.e., foundation model, AI Agent and Multi-agent collaboration).
3. Existing MAS frameworks usually design multiple agents with homogeneous traits, such as all being language-based agents. But when connecting MAS to real-world scenarios, it usually involves different kinds of AI agents. For example, we may need to bridge the connections between language-based agent, digital agent and robotic agents. However, these agents adopt various settings, from the inputs to the outputs. How to establish the connection between these agent is still a open problem that need to be handle in the future.